基于降噪自动编码器的中文新闻文本分类方法研究*

刘红光　马双刚　刘桂锋
(江苏大学科技信息研究所　镇江 212013)

摘要: 【目的】借助深度学习理论, 解决传统特征选择方法容易导致特征项不明确、分类精度下降的问题。【方法】对中文新闻文本进行分类时, 使用降噪自动编码器构建一个深层网络来学习对文本的压缩及分布式的表示, 并在网络最后一层采用 SVM 算法将其分类到具体的类别中去。【结果】随着样本数目的增大, 分类准确率、召回率和 F 值都在上升, 且比 KNN 算法、BP 算法和 SVM 算法取得了更优的分类效果, 平均分类准确率达到 95%以上。【局限】数据量依然较小, 且并没有完全发挥深度学习并行处理大容量数据的优势。【结论】该方法能提高特征项提取的准确性, 并能提高分类效果。

关键词: 降噪自动编码器　支持向量机　特征提取　文本分类
分类号: G350

1 引　言

信息技术的飞速发展, 使得海量的信息数据以指数级的模式不断增长, 标志着大数据时代的来临。在此背景下, 对海量文本信息的有效组织与利用显得尤为重要。文本分类技术以其对海量信息高效、准确地管理和定位的优势被广泛应用在社会生活的各个领域, 并取得了长足的发展。

在文本分类过程中, 一般采用向量空间模型(Vector Space Model, VSM)对文本进行表示。而文本数据结构和语义的复杂性, 使得经分词、删除停用词后的特征向量空间维度依然很高, 需要对其进一步优化。最常用的方法就是进行降维操作, 降维之后文本分类器要处理的文本数据规模大大降低, 噪声也大大减少。特征降维的常用方法有: 特征选择和特征提取。特征选择一般采用基于统计的方法, 得到的特征集是原始特征集中的一个子集, 常见的有卡方检验[1](CHI-Square) 、互信息[2](Mutual Information, MI)、信息增益[3](Information Gain, IG)等。关于特征选择方法的相关研究[4]表明: IG方法的性能相对较好。特征提取方法能够从原特征集中构造或者合成新的特征项, 从而降低文本特征的空间维度, 研究人员先后提出了许多不同的特征提取方法, 如互近邻聚类算法[5]、最大熵模型[6]等。虽然这些传统的特征选择或提取方法能识别出大部分特征, 但是也普遍存在着特征识别度较差的问题。如指定类别中很少出现但在其他类别中频繁出现的特征可能会被选择出来, 进而导致特征项丢失; 而经过提取后的特征可能会出现误差, 因而不能准确代表原有数据集, 尤其是从数据量较大、维数较多的数据集中提取出的特征项, 更容易出现误差, 最后导致分类精度下降。

2006年, Hinton等[7]介绍自动编码器(Auto Encoder, AE)构建的深度网络在图像和文本的特征降维方面的应用, 取得了比传统的特征降维方法更优的效果。因此, 学者纷纷将AE应用到特征提取过程中, 并不断提出稀疏自动编码器[8](Sparse Auto Encoder, SAE)、降噪

通讯作者: 马双刚, ORCID: 0000-0002-0957-898X, E-mail: 1107306870@qq.com。
*本文系教育部人文社会科学研究青年基金项目"基于超图模型的专利文本多标签分类研究"(项目编号:14YJC870014)的研究成果之一。

自动编码器[9](Denoising Auto Encoder, DAE)和卷积自动编码器[10](Convolutional Auto Encoder, CAE)等不同改进算法。其中，DAE在特征提取中的应用较为广泛，主要应用于对动态视频纹理[11]、音频[12]、图像[13]的特征提取中，在医学诊断[14]中也有所应用。本文只对DAE在文本特征提取中的应用进行深入研究。

文本中存在许多噪声，影响着分类的精度。因此相关学者选择采用DAE进行文本特征的提取，如刘勘等[15]针对短文本的特点，提出一种基于深层DAE的特征提取及聚类算法，有效地解决了短文本空间向量的高维、稀疏问题；秦胜君等[16]通过改进DAE，实现了无监督的样本分类，对不平衡率较高的样本具有良好的适应性。虽然这部分研究相对较少，但是可以看到，DAE构建的深度网络能够针对文本数据中噪声较大的特点，提取出更加准确地代表原始文本的特征编码，并有效去除其中的噪声，再结合分类算法进行文本分类时能够大大提高分类准确率。

与文献[15-16]不同，本文将DAE应用到新闻文本的特征提取中，首先使用DAE构建深度网络自动学习得到文本的低维特征；然后在网络的最顶层采用线性分类器支持向量机(Support Vector Machine, SVM)算法对得到的低维特征编码进行分类输出，根据输出的结果实现分类；最后分别与K近邻(K-Nearest Neighbors, KNN)算法、SVM算法和反向传播神经网络(Error Back Propagation, BP)算法进行比较，证明此方法的有效性。

2　相关理论基础

2.1　基于 DAE 的特征提取

AE[7, 17]构造的是一种无监督的深度网络结构，首先经过无监督的逐层贪心预训练与系统性的参数优化，得到多层非线性网络，然后利用此网络从无类标数据中提取出高维复杂输入数据的分层特征，并得到原始数据的分布式特征表示，能够比较好地复现输入的数据信号。AE 主要由两个部分组成：编码器和解码器，结构示意图如图 1 所示：

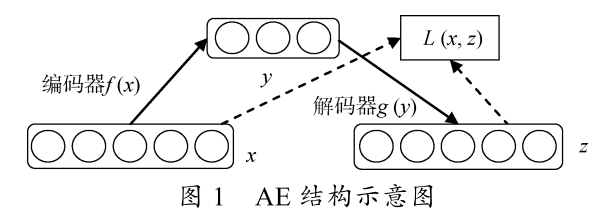


图 1　AE 结构示意图

但是，AE 无法消除数据中的噪声干扰。为了消除噪声干扰，获得更加鲁棒的特征，Vincent 等[18]提出可以用概率分布(通常使用二项分布)随机处理原始输入矩阵 X。对原始数据进行破坏处理得到 $\hat{X}$，然后对 $\hat{X}$ 进行编码处理，后续过程即与 AE 的运算过程相同，此改进后的编码器即为降噪自动编码器，结构示意图如图 2 所示：

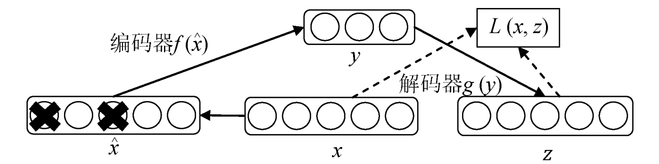


图 2　降噪自动编码器结构模型示意图

编码器 $f(\hat{x})$ 用于高维数据的降维，首先对输入向量 x 进行破坏处理得到 $\hat{x}$，然后输入到编码器 $f(\hat{x})$，经过线性变换和激活函数的作用，最后得到隐含的编码结果y。解码器g(y)用于低维编码的重构过程，即将隐含层数据映射回重构z，分别表示为如下函数：

$$y = f(\hat{x}) = S_f(W\hat{x} + b_y) \quad (1)$$

$$z = g(y) = S_g(W'y + b_z) \quad (2)$$

其中，S_f是非线性激活函数，其表达式为：

$$S_f = \text{sigmoid}(y) = \frac{1}{1+e^{-y}} \quad (3)$$

S_g 是解码器的激活函数，本文也采用 sigmoid 函数，$W' = W^T$，是 W 的转置，因此只需要训练 W 即可，b_y 和 b_z 是偏倚向量。

DAE 的训练过程即是在训练样本集 D 上寻找参数 $\theta = \{W, b_y, b_z\}$ 的最小化重构误差，重构误差的表达式如下：

$$J_{AE} = \sum_{x \in D} L(x, g(f(\hat{x}))) \quad (4)$$

其中，L 为重构误差函数，文献[19]表明在实验过程中，交叉熵损失函数一直优于平方差损失函数，因此本文采用交叉熵损失函数，表达式如下：

$$L(x,z) = -\frac{1}{n}\sum_{i=1}^{d_x}[x_i \ln z_i + (1-x_i)\ln(1-z_i)] \quad (5)$$

其中，n 是训练集样本数，x_i 是第 i 个输入，z_i 为对应的第 i 个解码重构后的数据。

自动编码器采用经典的随机梯度下降算法进行训练，在每个迭代过程中，利用公式(6)更新权重矩阵：

$$W \leftarrow W - \varphi \times \frac{\partial L(x,y)}{\partial W}$$

$$b_y \leftarrow b_y - \varphi \times \frac{\partial L(x,y)}{\partial b_y} \quad b_z \leftarrow b_z - \varphi \times \frac{\partial L(x,y)}{\partial b_z} \tag{6}$$

其中，φ是学习率，b_y和b_z采用与之相同的更新方式。

2.2 基于SVM的分类

SVM算法[20]的训练过程是要找到一个超平面，使得这个超平面的正反例分别落在两侧，在所有超平面中与正反例的距离最大且到最近的正反例的距离相等，然后对未知类别的样本数据，计算其位于超平面一侧，即为其分属的类别。

在线性可分的情况下，分类线性方程为$(w \cdot x) + b = 0$，对此方程进行正则化，使得每一个线性可分的样本$(x_i, y_i), i = 1,2,\cdots,l, x \in R^n, y \in \{-1,+1\}$，均满足:

$$y_i[(w \cdot x_i) + b] - 1 \geqslant 0 \quad i = 1,2,\cdots,l \tag{7}$$

其中: x_i是输入的第i个样本，l为样本数，w是可调的权值向量，b是偏置。$y_i \in \{-1,+1\}$表示相应x_i的期望分类。

为了求得最优分类超平面，需要在满足公式(7)下使得分类间隔$margin = 2/\|w\|$最大，即使得$\|w\|^2$最小，这是一个典型的二次规划问题，目标函数为:

$$\min_w L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l}\alpha_i\{[(w \cdot x_i) + b] - 1\} \tag{8}$$

利用拉格朗日优化方法可以将上述问题转化为其对偶问题，即加入约束条件$\sum_{i=1}^{l}\alpha_i y_i = 0$和$\alpha_i \geqslant 0, i = 1,2,\cdots l$，对$\alpha_i$求解下列函数的极大值:

$$Q(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j \cdot y_i y_j (x_i \cdot y_i) \tag{9}$$

$\alpha_i \geqslant 0$为与每个样本相对应的拉格朗日稀疏，即训练样本中仅有少数的拉格朗日系数α_i^*不为0，这样的样本定义为支持向量。

在最优分类面中采用适当的核函数就可以实现某一非线性变换后的线性分类，而计算的复杂度却没有增加。此时的目标函数公式(9)变为:

$$Q(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j \cdot y_i y_j K(x_i, y_i) \tag{10}$$

最后训练后的相应的分类函数为:

$$d(x) = \sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b^* \tag{11}$$

即支持向量机，根据d(x)的符号来确定输入样本x的归属。

3 DAS文本分类模型

本文设计的结合DAE和SVM算法的中文新闻文本分类模型(简称DAS分类模型)，主要包括6个部分，即NLPIR文本分词、去除停用词、文本表示、DAE特征提取、SVM分类和分类效果评价，如图3所示:

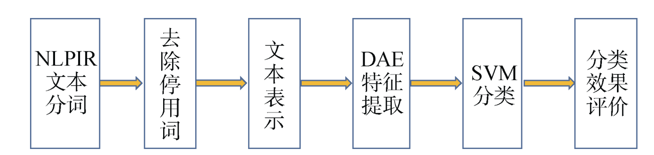


图3 中文新闻文本分类模型示意图

(1) 中文不同于西文，词与词之间没有明显的分割界限，因此需要对中文文本进行特殊的分词操作。本文采用比较成熟的NLPIR汉语分词系统[21]对中文新闻文本进行分词操作。

(2) 经过步骤(1)分词后形成的词语有大量的停用词，包括标点符号和一些对分类不起作用的常见词等，本文收集多个停用词表后合并成一个较全面的停用词表，用来剔除这些停用词，得到能代表文本特征的候选特征词。

(3) 经过步骤(2)得到的候选特征词依然很多，维数特别大，需要对其进行初步筛选，本文通过信息增益算法对文本特征进行初步筛选后，采用VSM模型进行文本的特征表示。

(4) 将经过步骤(3)得到的特征表示输入一个由DAE构建的深度网络中，经过逐层训练后，得到一个维数比较低的特征编码。

(5) 在深度网络的最后一层，用SVM算法对经过步骤(4)得到的特征编码进行分类输出，根据输出结果进行分类。

(6) 对分类的效果进行评价，并根据评价结果不断地对此文本分类模型进行优化，直至得到满意的分类结果。

其中，文本分类中最基础最重要的工作是步骤(4)特征词的提取，而文本中存在大量的冗余数据和噪声，在提取特征词的时候容易导致误差的产生和识别度较

差的问题，进而影响最终的分类效果。要想取得比较好的分类效果，需要将这些冗余数据或噪声的影响尽可能降低到最小。DAE 将输入数据进行破坏处理后，利用这些破损数据训练出来的特征系数噪声比较小，并且破损数据在一定程度上能够减轻训练数据与测试数据的代沟。

因此，为了从文本中提取、编码出更加鲁棒的特征并消除噪声的影响，以取得较好的分类效果，本文借鉴相关理论[7, 17-20]，将 DAE 应用到中文新闻文本特征词的提取中，构建一个深度网络，逐层训练得到一个低维的特征编码，提取出最能代表文本的低维特征，实现高维文本数据的特征降维过程，并利用 SVM 算法在深度网络的最顶层对输出的低维编码进行分类输出，根据输出的结果实现最终的分类过程。基于 DAE 和 SVM 构建的深度网络如图 4 所示：

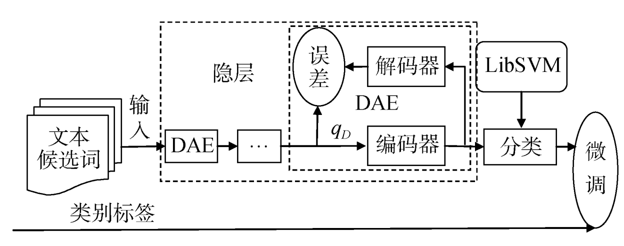


图 4　分类训练过程

其中，文本候选词首先经过由多层DAE构成的隐层处理后，得到低维编码，在最顶层由LibSVM对低维编码进行分类输出，根据分类输出的结果进行微调，实现整个文本分类模型的训练过程。

4　仿真实验

4.1　仿真实验步骤

(1) 实验 1：经典实验

为了测试本文 DAS 分类模型的优越性，在相同的数据集上，采用信息增益的特征选择方法选择出特征后，采用经典的训练算法，分别为 KNN 算法、SVM 算法和只包含一层隐藏神经元的 BP 神经网络算法，进行分类仿真实验，并将其分类召回率、准确率和 F 值与本文 DAS 分类模型进行比较。采用经典算法作对比仿真实验的具体步骤如下：

①选择仿真实验数据集。仿真实验的新闻文本数据集[22]由复旦大学计算机信息与技术系李荣陆提供，数据集标注比较规范，规模适中，适合中小型的分类仿真实验。

此数据集中 answer 分组为测试语料，共 9 833 篇文档，train 分组为训练语料，共 9 804 篇文档，分为 20 个类别。随机选取 6 个类别，每个类别 1 000 篇，分别以 200 篇、400 篇、600 篇、800 篇设置 4 组训练集，其中每组都以 200 篇作为测试集，分别进行训练。具体类别信息及实验分组设计如表 1 所示：

表 1　文本分类实验具体类别信息及分组设计

类别 / 分组	类别名	训练集(4 组)	测试集
C01	Computer	200、400、600、800	200
C02	Environment	200、400、600、800	200
C03	Agriculture	200、400、600、800	200
C04	Economy	200、400、600、800	200
C05	Politics	200、400、600、800	200
C06	Sports	200、400、600、800	200

②文本数据集的预处理包括文本分词和去除停用词。文本分词采用的 NLPIR 汉语分词系统，其主要功能包括中文分词、词性标注、命名实体识别、用户词典功能、微博分词、新词发现与关键词提取等，是国内比较成熟、用户较多的中文文本分词系统。本文对文本语义特征进行分析，并综合网络上的停用词表，制作了一个比较全面的停用词表，如表 2 所示：

表 2　停用词表

标点符号	特殊符号	无意义词	数字	西文字符
。	<	的	1	A(a)
、	>	啊	2	B(b)
.	/	一个	3	C(c)
:	@	你	4	D(d)
,	~	本文	5	E(e)
…	…	…	…	…

③文本表示。经过预处理后的文本维数过大，需要对其进行初步的降维处理，计算每个特征词的信息增益值，公式如下：

$$IG_{(t)} = -\sum_{i=1}^{m} P_{(c_i)} \log P_{(c_i)} + P_{(t)} \sum_{i=1}^{m} P_{(c_i|t)} \log P_{(c_i|t)} + P_{(\bar{t})} \sum_{i=1}^{m} P_{(c_i|\bar{t})} \log P_{(c_i|\bar{t})} \quad (12)$$

其中，m 为总类别数，c_i 代表类别，$P_{(c_i)}$ 为类别 c_i 出现的概率；$P_{(t)}$ 为包含特征词的文档的概率，$P_{(\bar{t})}$ 为不包含特征词的文档的概率；$P_{(c_i|t)}$ 为包含特征 t 属于 c_i 的概率；$P_{(c_i|\bar{t})}$ 为包含特征 t 但属于 c_i 的概率。

计算出信息增益值后，将其按大小排序并保留前 5 000 个特征词用向量空间模型表示，如表 3 所示。

表 3　文本特征集的向量空间模型表示

文本＼特征词	t_1	…	t_j	…	t_n
d_1	w_{11}	…	w_{1j}	…	w_{1n}
…	…	…	…	…	…
d_i	w_{i1}	…	w_{ij}	…	w_{in}
…	…	…	…	…	…
d_m	w_{m1}	…	w_{mj}	…	w_{mn}

在该特征词矩阵中，n 表示所有文档中的特征词总数，每个特征词对应特征空间中的一维；m 代表所有待分类的文本数；将每个文档表示成 N 维空间中的一点，如：$V(d_i)=((t_1,w_{i1}),(t_2,w_{i2}),\cdots,(t_k,w_{ik}),\cdots,(t_n,w_{in}))$，特征权重值 w_{ij} 为每个特征词的 TF-IDF 值，计算公式如下：

$$W_{(t,d)}=\frac{tf_{(t,d)}\times\log(N/n_t+a)}{\sqrt{\sum_{t\in d}[tf_{(t,d)}\times\log(N/n_t+a)]^2}} \tag{13}$$

其中，$W_{(t,d)}$为特征词 t 在文本 d 中的权重；$tf_{(t,d)}$表示词条 t 在文档 d 中出现的频数；n_t 为文本集中含有特征 t 的文本的数量；$\log(N/n_t+a)$为逆文本频率函数，n_t 越大此值越小，a 为一个常量，本文取 0.01；分母是一个归一化因子。

④分类训练。利用分类算法进行有监督的分类训练，得到分类参数，并用测试数据集进行分类测试。本文选择的算法分别为：KNN 算法，此算法相对比较简单，用 C 语言自主设计的 KNN 算法；SVM 算法，采用比较成熟的 LibSVM 进行分类实验；BP 算法，MATLAB 自带的成熟的神经网络工具箱。

⑤分类效果评价与比较。采用召回率 R(Recall)、准确率 P(Precision)和 F 值对最终的分类结果进行评价。公式如下：

$$R=\frac{M}{M+T} \tag{14}$$

$$P=\frac{M}{M+N} \tag{15}$$

$$F=\frac{2\times R\times P}{R+P} \tag{16}$$

其中，M 为正确分类到该类的文本数，N 为错分到该类中的文本数，T 为属于该类却错分为别类的文本数。

(2) 实验 2：优化实验

本文设计的 DAS 分类模型用于新闻文本分类的仿真实验的步骤①–步骤③和步骤⑤都与经典实验完全相同，只有步骤④分类训练与经典实验步骤不同。

在本文设计的 DAS 分类模型中，将训练文本数据集经过步骤①–步骤③，得到矩阵表示后，输入一个 DAE 构建的深度网络，用非监督学习方法对 5 000 维的特征进行逐层的降维操作，并在最后一层采用线性分类器 SVM 算法对文本进行分类输出，根据输出结果调整训练过程中的各个参数，得到最终分类参数后利用测试数据集进行测试。

①特征降维。将文本的向量化矩阵表示 X 先经过破坏处理得到矩阵 $\hat{X}$，然后将破坏后的矩阵 $\hat{X}$ 输入到编码器得到编码，再经过解码器重构得到一个重构矩阵，将重构矩阵与原始矩阵比较得到重构误差，调整编码器和解码器的参数使得重构误差最小，得到最终的编码。将上层中得到的编码特征作为下层的输入，采用相同的方法得到下层的编码，如此不断进行，得到规定数量层数的编码。

本文设计的深层次网络的节点数分别为 5000-2500-1200-600-300-100-50-20，加上最终的线性分类器 SVM 共 9 层，先对每层的矩阵表示经过一个随机化置 0 的过程，再进行训练，每层训练结束后继续训练下一层，直到完成降噪自编码器的降维过程。

②有监督微调。将步骤①得到的 20 维的特征编码，应用 SVM 算法对其进行分类输出，然后根据输出的结果分类。再对各层的系数进行微调。此监督训练完成后，即用来对测试集的文本进行分类，以测试这个分类系统的有效性。

本文采用 LibSVM 算法对每个文本降维后获取到的特征编码进行分类，然后用 BP 算法从顶层向下进行各层系数的微调，最终取得调整后的系数，即可用来对测试数据集进行测试。

4.2　实验结果与分析

在 4 组训练集下分别进行实验，得出分类召回率、准确率与F值的情况如图 5 所示：

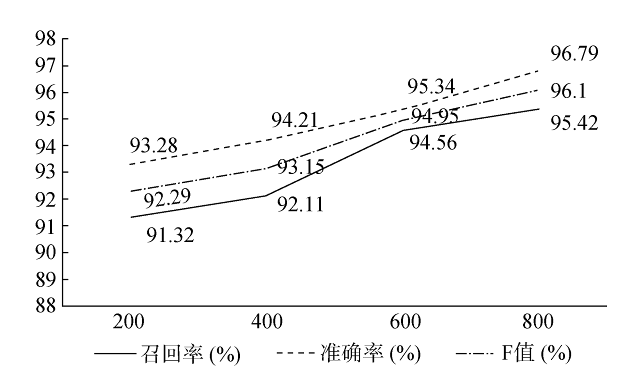


图 5　分类召回率、准确率与 F 值随数据集变化的曲线图

可以看到，DAS分类模型随着训练集的增大，分类召回率、准确率和F值都在不断增大。这是由于深度网络对数据集的数量要求比较高，过小的数据集会导致产生过拟合现象从而导致分类效果欠佳。因此，针对深度网络，数据集的大小能够决定网络训练的效果，数据集越大，越能训练得到较好的分类效果。

经过与不同分类算法比较的仿真实验，得出4组实验下各算法的分类召回率、准确率和F值如表4至表6所示：

表 4　不同训练集下各算法的分类召回率(%)

训练集＼算法	KNN	BP	SVM	DAS
200	83.32	87.38	94.18	91.32
400	84.54	89.57	93.73	92.11
600	83.97	91.23	93.66	94.56
800	84.21	93.85	93.89	95.42

表 5　不同训练集下各算法的分类准确率(%)

训练集＼算法	KNN	BP	SVM	DAS
200	87.68	90.32	93.01	93.28
400	88.72	90.34	94.71	94.21
600	86.35	92.15	93.78	95.34
800	85.78	94.24	94.59	96.79

表 6　不同训练集下各算法的分类 F 值(%)

训练集＼算法	KNN	BP	SVM	DAS
200	85.44	88.83	93.59	92.29
400	86.58	89.95	94.22	93.15
600	85.14	91.69	93.72	94.95
800	84.99	94.04	94.24	96.10

KNN算法在训练过程中，生成所有训练文本的特征向量，在测试过程中，比较测试文本的特征向量与所有训练文本特征向量的相似度，在中小型的分类实验中能够取得不错的效果。但是可以看到这种方法在很大程度上依赖于选出的特征词，如果选出的特征词代表性不强，分类效果会变得比较差，也可以看到，KNN算法取得的分类效果比其他算法差；BP算法是一种典型的浅层神经网络算法，能够反向传播误差，将误差分摊给各层单元，进而修正各单元的权值系数，完成训练过程。但是BP算法往往只设计三层的网络，在训练集比较少的情况下分类效果较差，随着训练集数目的增加，分类召回率、准确率和F值都有不同程度的增加；SVM算法的训练过程是要找到一个超平面，使得这个超平面的正反例分别落在两侧，在所有超平面中与正反例的距离最大且到最近的正反例的距离相等，然后对未知类别的文本，计算其位于超平面的一侧，即为其分属的类别，在对小规模样本的数据集的处理中颇占优势，但是对大样本数据集的分类效果略差，随着数据集的增大，分类效果并没有明显提升；DAS分类模型利用降噪自动编码器无监督地学习到新闻文本的特征编码，符合人脑以词–句–段–意的逐层分析方式对文本的理解，能够更精确地模拟人脑对文本所表达的意思的理解过程，因此能够取得更好的分类效果，从表4至表6也可以看出，本文的DAS分类模型分类效果更好。

5　结　语

深度学习已经在学术界、工业界掀起了研究热潮，并取得了相当大的成果。本文借鉴 DAE 和 SVM 的相关理论，设计DAS分类模型，将DAE构建的深度网络应用于中文新闻文本的特征降维过程中，并在深度网络的最顶层用 SVM 进行分类，根据分类的结果不断微调各层的系数，最终用测试数据集测试分类效果。结果表明，DAS 分类模型降低了新闻文本数据中噪声的影响，分类效果比较好，能够取得比 KNN、SVM 和 BP 算法更好的分类效果。但是也看到，虽然设置了 4 组不同数据集的仿真实验，然而数据量依然比较小，并没有完全发挥深度学习并行处理大容量数据的优势，下一步的研究工作将集中在对大数据集的实验上。

参考文献：

[1] 裴英博，刘晓霞．文本分类中改进型 CHI 特征选择方法的研究[J]．计算机工程与应用，2011，47(4)：128-130．(Pei Yingbo, Liu Xiaoxia. Study on Improved CHI for Feature Selection in Chinese Text Categorization [J]. Computer Engineering and Applications, 2011, 47(4): 128-130.)

[2] 辛竹，周亚建．文本分类中互信息特征选择方法的研究与算法改进[J]．计算机应用，2013，33(S2)：116-118，152．(Xin Zhu, Zhou Yajian. Study and Improvement of Mutual Information for Feature Selection in Text Categorization [J]. Journal of Computer Applications, 2013, 33(S2): 116-118, 152.)

[3] 郭颂，马飞．文本分类中信息增益特征选择算法的改进[J]．计算机应用与软件，2013，30(8)：139-142．(Guo Song, Ma Fei. Improving the Algorithm of Information Gain Feature Selection in Text Classification [J]. Computer Applications and Software, 2013, 30(8): 139-142.)

[4] Peters C, Koster C H. Uncertainty-based Noise Reduction and Term Selection in Text Categorization [C]. In: Proceedings of

the 24th BCS-IRSG European Colloquium on IR Research, Glasgow, UK. Springer, 2002: 248-267.

[5] Lewis D D. Representation and Learning in Information Retrieval [D]. University of Massachusetts, 1992.

[6] 李学相. 改进的最大熵权值算法在文本分类中的应用[J]. 计算机科学, 2012, 39(6): 210-212. (Li Xuexiang. Research of Text Categorization Based on Improved Maximum Entropy Algorithm [J]. Computer Science, 2012, 39(6): 210-212.)

[7] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks [J]. Science, 2006, 313(5786): 504-507.

[8] Bengio Y, Lamblin P, Popovici D, et al. Greedy Layer-wise Training of Deep Networks [C]. In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada. 2007, 19: 153.

[9] Vincent P, Larochelle H, Bengio Y, et al. Extracting and Composing Robust Features with Denoising Autoencoders [C]. In: Proceedings of the 25th International Conference on Machine Learning. ACM, 2008: 1096-1103.

[10] Masci J, Meier U, Cireşan D, et al. Stacked Convolutional Auto-encoders for Hierarchical Feature Extraction [C]. In: Proceedings of the 21st International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, 2011: 52-59.

[11] 汪彩霞, 魏雪云, 王彪. 基于堆栈降噪自动编码模型的动态纹理分类方法[J]. 现代电子技术, 2015, 38(6): 20-24. (Wang Caixia, Wei Xueyun, Wang Biao. Dynamic Texture Classification Method Based on Stacked Denoising Autoencoding Model [J]. Modern Electronics Technique, 2015, 38(6): 20-24.)

[12] Wu Z, Takaki S, Yamagishi J. Deep Denoising Auto-encoder for Statistical Speech Synthesis [OL]. arXiv: 1506.05268, 2015.

[13] Li J, Struzik Z, Zhang L, et al. Feature Learning from Incomplete EEG with Denoising Autoencoder [J]. Neurocomputing, 2015, 165: 23-31.

[14] 胡帅, 袁志勇, 肖玲, 等. 基于改进的多层降噪自编码算法临床分类诊断研究[J]. 计算机应用研究, 2015, 32(5): 1417-1420. (Hu Shuai, Yuan Zhiyong, Xiao Ling, et al. Stacked Denoising Autoencoders Applied to Clinical Diagnose and Classification [J]. Application Research of Computers, 2015, 32(5): 1417-1420.)

[15] 刘勘, 袁蕴英. 基于自动编码器的短文本特征提取及聚类研究[J]. 北京大学学报: 自然科学版, 2015, 51(2): 282-288. (Liu Kan, Yuan Yunying. Short Texts Feature Extraction and Clustering Based on Auto-Encoder [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2015, 51(2): 282-288.)

[16] 秦胜君, 卢志平. 基于降噪自动编码器的不平衡情感分类研究[J]. 科学技术与工程, 2014, 14(12): 232-235. (Qin Shengjun, Lu Zhiping. Research of Unbalance Sentiment Classification Based on Denoising Autoencoders [J]. Science Technology and Engineering, 2014, 51(12): 232-235.)

[17] Bengio Y, Delalleau O. On the Expressive Power of Deep Architectures [C]. In: Proceedings of the 22nd International Conference on Algorithmic Learning Theory. Springer Berlin Heidelberg, 2011: 18-36.

[18] Vincent P, Larochelle H, Bengio Y, et al. Extracting and Composing Robust Features with Denoising Autoencoders [C]. In: Proceedings of the 25th International Conference on Machine Learning. ACM, 2008: 1096-1103.

[19] Neural Networks and Deep Learning [EB/OL]. [2015-12-23]. http://neuralnetworksanddeeplearning.com/chap3.html.

[20] Vapnik V N. The Nature of Statistical Learning Theory[J]. IEEE Transactions on Neural Networks, 1995, 10(5): 988-999.

[21] NLPIR 汉语分词系统[EB/OL]. [2015-09-22]. http://ictclas.nlpir.org/. (NLPIR Chinese Word Segmentation System [EB/OL]. [2015-09-22]. http://ictclas.nlpir.org/.)

[22] 文本分类语料库(复旦)测试语料 [EB/OL]. [2015-12-24]. http://www.nlpir.org/?action-viewnews-itemid-103. (Text Categorization Corpus (Fudan) Test Corpus [EB/OL]. [2015-12-24]. http://www.nlpir.org/?action-viewnews-itemid-103.)

作者贡献声明:

刘红光: 提出研究思路, 设计研究方案;
马双刚: 采集、清洗和分析数据, 选择合适的算法进行实验;
刘桂锋: 验证实验可行性, 整理实验数据, 并对实验进行总结。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 刘红光, 马双刚, 刘桂锋. 采用的实验数据集.rar. 从文本分类语料库测试语料中选择的 6 个类别数据集.

[2] 刘红光, 马双刚, 刘桂锋. InfoGain.txt. 由信息增益算法筛选出来的特征词.

[3] 刘红光, 马双刚, 刘桂锋. TF-IDF 值.rar. 计算出的各个文本的 TF-IDF 值.

[4] 刘红光, 马双刚, 刘桂锋. 分类结果集.xlsx. 各个算法计算的分类结果.

收稿日期: 2016-01-13
收修改稿日期: 2016-02-19

Classifying Chinese News Texts with Denoising Auto Encoder

Liu Hongguang　Ma Shuanggang　Liu Guifeng
(Institute of Scientific & Technical Information, Jiangsu University, Zhenjiang 212013, China)

Abstract: [**Objective**] This paper proposes a new method to improve the classification accuracy of the Chinese news texts with the help of Deep Learning theory. [**Methods**] We first used the denoising auto encoder to construct a deep network to learn the zipped and distributed representation of the Chinese news texts. Second, we used the SVM algorithm to classify these news texts. [**Results**] As the number of samples expanding, the precision rate, the recall rate and the F value of the proposed method increased too. The results are better than those of the applications using the KNN, BP and SVM algorithms. The average precision rate was higher than 95%. [**Limitations**] The data size was relatively small, thus, the proposed method did not fully utilize the parallel data processing capacity of the deep learning technology. [**Conclusions**] The proposed method improves the performance of applications classifying Chinese news texts.

Keywords: DAE　SVM　Feature selection　Document classification